Eötvös Loránd Tudományegyetem

Bölcsészettudományi Kar

DOKTORI DISSZERTÁCIÓ

BESZÉLŐDETEKTÁLÁS MAGYAR NYELVŰ SPONTÁN TÁRSALGÁSOKBAN

BEKE ANDRÁS

Nyelvtudományi Doktori Iskola
vezető: Prof. Dr. Bárdosi Vilmos egyetemi tanár

Alkalmazott Nyelvészet Doktori Program
vezető: Prof. Dr. Gósy Mária egyetemi tanár

A bizottság tagjai

A bizottság elnöke:	Prof. Dr. Nyomárkay István akadémikus
Hivatalosan felkért bírálók:	Prof. Dr. Adamikné Jászó Anna DSc
	Prof. Dr. Olaszy Gábor DSc
A bizottság titkára:	Dr. Dér Csilla PhD
A bizottság további tagjai:	Prof. Dr. Vicsi Klára DSc
	Prof. Dr. Balázs Géza CSc
	Dr. Hámori Ágnes PhD

Témavezető
Prof. Dr. Gósy Mária DSc

Budapest, 2014

1. BEVEZETÉS

A konverzációelemzés adta keretben a társalgás strukturális módon épül fel (Garfinkel 1967; Goffman 1983; Schegloff 1992; Sacks et al. 1974; Sacks 1992; Iványi 2001; Stokoe 2006). Ha a társalgás rendszerszerű, akkor feltételezhetően gépi úton modellezhető, tehát vizsgálata nemcsak nyelvészeti, hanem beszédtechnológiai szempontból is fontos. A beszédtechnológia „*mesterséges intelligencián belül a beszéd alapú (verbális) gyakorlati alkalmazások kifejlesztésével és létrehozásával foglalkozik*" (Németh–Olaszy 2010). Az ember-gép verbális kommunikációban számos részfeladatot modelleztek már magyar nyelven, mint a beszéd gépi megértését (beszédfelismerés), illetve a gépi beszédelőállítást (beszédszintézis), a beszélő személy gépi azonosítását a hangja alapján (beszélőfelismerés). Ezen részfolyamatok a társalgásban kapcsolódnak össze, ahol nemcsak egyoldalú a folyamat, vagyis nemcsak beszédfelismerésről vagy beszédelőállításról beszélhetünk, hanem ezek körkörös folyamatáról, ami a beszélők váltakozásából fakad, vagyis fontos lépés, hogy ezt a folyamatot gépileg tudjuk lekövetni, előjelezni. Ezt a dinamikus rendszert, amelyben a beszélők váltakozását igyekeznek leírni gépi eszközökkel, a beszédtechnológiában beszélődetektálásnak (speaker diarization) nevezik. A beszélődetektálás során a folyamatos beszédben az akusztikai jelből gépi úton határozzuk meg, hogy mikor ki beszél (Jin et al. 2004). A beszélődetektálás során a folyamatos társalgásokat automatikusan beszélőkre szegmentáljuk, így a társalgások szövegeit beszélőkhöz rendelhetjük, így a szöveg sokkal könnyebben feldolgozható más, például tartalomkinyerő algoritmusok számára.

A beszélődetektálást alapvetően két alfeladatra lehet bontani (Jin et al. 2004; Kotti et al. 2008): a beszélő szerinti szegmentálásra (speaker segmentation) és a beszélőosztályozásra (speaker clustering). Az első feladat célja elkülöníteni az azonos beszélőktől származó beszédrészeket, a második részfolyamatban pedig ezeket a szegmentumokat kell osztályozni a beszélők szerint. A két részfeladaton kívül fontos feladat még a beszéddetektálás és az egyszerre beszélések detektálása.

A beszélődetektálás megvalósítására jelentős mennyiségű kutatás történt idegen nyelvre (Tritschler–Gopinath 1999; Sivakumaran–Fortuna–Ariyaeeinia 2001; Lu–Zhang 2002a; Cettolo–Vescovi 2003; Shih-Sian Cheng et al. 2010; Vescovi–Cettolo–Rizzi 2003). Magyar nyelvre azonban idáig nem született olyan munka, amely a beszélődetektálás megvalósítását tűzte volna ki céljául. A jelen értekezés célja, hogy első ízben hozzon létre magyar spontán társalgásokra működő beszélődetektáló rendszert. A dolgozat célkitűzése egyrészt az, hogy a beszélődetektáláshoz kapcsolódó tudományterületeket bemutassa, illetve hogy maga a beszélődetektálás főbb módszertani ismeretit leírja. A másik célja az volt, hogy a beszélődetektálóhoz szükséges algoritmusokat elkészítsük (egyszerre beszélés-detektálás, beszélőszegmentáló, beszélőklaszterező), és a már létező algoritmusokat implementáljuk a beszélődetektálóba (beszéd/nem beszéddetektáló, beszélőfelismerő algoritmus). Az általunk javasolt rendszer célja, hogy magyar nyelvű spontán

társalgásokban automatikusan detektálja a beszélőket pusztán akusztikai információk alapján: vagyis megoldást adjon arra a kérdésre, hogy „*Mikor ki beszél?”*. Az általunk javasolt beszélődetektáló rendszer alapvetően nem-ellenőrzött tanulási eljárásokon alapul.

2. AZ ÉRTEKEZÉS FELÉPÍTÉSE

A disszertáció 11 fejezetből áll. Az első legnagyobb fejezetben bemutatjuk a témához kapcsolódó tudományágak elméleti és gyakorlati hátterét. A beszédprodukció és beszédpercepció ismertetésén keresztül bemutatjuk a spontán beszéd jellegzetességeit, majd annak a legtöbbet használt változatát, a társalgást. Ebben a fejezetben prezentáljuk a beszélőalkalmazkodás részletezésével a dinamikusan változó társalgást. Szintén ebben a fejezetben kap helyet a társalgás építőköveinek bemutatása, amely a beszédforduló (angol terminussal: turn). A Bevezetés fejezetben mutatjuk be az egyik lehetséges jelzési eszközt, amely a diskurzusjelölő. Az 1. fejezetben taglaljuk az egyszerre beszélések szerepét a társalgásokban, illetve, hogy hogyan dolgozhatók fel a beszéddetektálás során.

A disszertáció 2. fejezete módszertani áttekintést ad a beszélődetektálásban használt algoritmusokról. Itt kerül bemutatásra a beszéddetektálás folyamata, amelynek célja, hogy a folyamatos akusztikai jelben jelölje, hogy hol van beszédrész, illetve nem beszédrész. A 2. fejezet ismerteti a beszélőfelismerés alapvető módszertani kérdéseit, amelynek célja, hogy milyen módon lehet gépileg felismerni a beszélő személyt a hangja alapján. Ebben a fejezetben kap helyet az egyszerre beszélések automatikus osztályozása is, amelynek igen nagy szerepe van a beszélődetektálás téves riasztásaink csökkentésében. A Módszertani fejezetben (2. fejezet) ismertetjük magát a beszélődetektáló rendszerek elméletét és módszertanát.

A saját kutatásunk céljainak, kérdéseinek és hipotézisének ismertetése a 3. fejezetben történik.

A 4. fejezetben a kísérleti személyek, általános anyag és módszer ismertetése történik. Itt mutatjuk be a kísérletekhez használt adatbázis felépítését, tartalmát, illetve itt kerül bemutatásra a beszélődetektálás kiértékeléséhez használt NIST (National Institute of Standards and Technology, Nemzetközi, Szabványok és Technológiák Nemzetközi Intézete) által javasolt DER (Detection Error Rate) eljárás, és az osztályozás kiértékeléséhez használt DET (Detection Error Tradeoff) algoritmus.

Az 5. fejezetben a kísérletek és az eredmények ismertetésére kerül sor. Először a beszéddetektálás folyamat lépéseit és eredményeit írjuk le. Ezután a beszélőspecifikus akusztikai jellemzők vizsgálatát és az azzal elért eredményeket prezentáljuk. A következő alfejezet a beszélődetektálásban legtöbb hibát okozó egyszerre beszéléseket detektáló rendszert mutatjuk be. Az 5. fejezet utolsó fejezetében pedig az általunk fejlesztett beszélődetektáló rendszert és az azzal elért eredményeket prezentáljuk.

A 6. fejezet az általános következtetéseket tartalmazza, amelyet az általános összefoglalás követ (7. fejezet). Ezután ismertetjük a beszélődetektálás felhasználási

és további fejlesztési lehetőségeit (8. fejezet). Végül a disszertáció téziseit fogalmazzuk meg (9. fejezet). Ezt követi az Irodalom (10), és a Rövidítések jegyzéke fejezet (11. fejezet).

3. KÍSÉRETI SZEMÉLYEK ÉS ANYAG

A disszertációban a BEA adatbázisból (Gósy 2012) 100 társalgást választottunk ki, amely 55 órányi társalgást jelent. A társalgásokban minden esetben három személy vett rész. Ebből két társalgó állandó volt (2 nő, életkoruk 32 év). A harmadik személy 43 férfi és 67 nő közül került ki, átlagos életkoruk 35 év.

A felvétel minősége laboratóriumi körülményekhez hasonló. A felvételt egy Audio-Technica AT 4040 típusú mikrofonnal, egy csatornára rögzítették 44 KHz-en, amelyet újramintavételeztünk 16 KHz-en.

A társalgások annotációi a következőket tartalmazták:

(i) Szünetek: minden olyan szünetet jelöltünk, amely meghaladta a 100 ms-ot. Nyilvánvalóan az artikulációból adódó némafázisokat nem jelöltük még akkor sem, ha azok ezen küszöböt átlépték is.

(ii) Beszélőváltások: folyamatos jelben bejelöltük, hogy mely időpillanatban van beszélőváltás, illetve hogy az egyes beszédszegmensek mely beszélőhöz tartoznak. A háttércsatorna-jelzéseket nem vettük beszédváltásnak, csak abban az esetben, ha tényleges szóátvételről volt szó.

(iii) Egyszerre beszélések: jelölve voltak azon részei is a beszédnek, ahol egy időben kettő vagy három személy szólalt meg. Nem jelöltük azonban azon részeket, ahol az átfedő beszéd nem haladta meg az 50 ms-ot, mivel ezek detektálása alapvetően nem megvalósítható.

4. EREDMÉNYEK

4.1 BESZÉDDETEKTÁLÓ

A beszélődetektálás kialakításban fontos szerepet játszik a beszéddetektálás (voice activity detection, VAD) megvalósítása, amelynek lényege, hogy automatikusan meghatározzuk az egyes jelszegmensekre, hogy az beszéd vagy nem-beszéd szegmens. A Giannakopoulos (2009) által kidolgozott és MATLAB-ba implementált beszéddetektáló algoritmusát használtuk, illetve módosítottuk. Ez az algoritmus rövid idejű energiafüggvény (short-term energy) és spektrális centroid (spectral centroid) akusztikai jellemzőket és adaptív küszöbölést alkalmaz a beszéd és nem-beszéd szegmensek automatikus meghatározására. Az általunk ajánlott módszer annyiban tér el ettől, hogy a küszöb meghatározását nem-ellenőrzött tanulási metódussal végezzük el (vö. (Ying et al. 2011). A középpontok (beszéd és nem-beszéd) megtalálásához klaszteranalízist használtunk, azon belül pedig a k-közép (k-means) algoritmust. A jelen kutatás célja annak tesztelése, hogy az általunk javasolt nem-ellenőrzött tanulási

módszer javít-e az eredményeken. A VAD kiértékeléséhez a DER (diarization error rate) módszert használtuk (NIST MD-eval-v12 DER kiértékelő scriptje 2006). Az alaprendszer és az általunk javasolt rendszer összehasonlításához nem-parametrikus összetartozó mintás (Wilcoxon-próba) tesztet használtunk, Monte Carlo szimulációval megerősítve.

A társalgásokban manuálisan jelöltük azokat a részeket, ahol valamelyik adatközlő beszél, illetve azokat a részeket, ahol nincs beszédjel, vagyis némaszünet van. A korpusz 49 órányi beszédrészt és 6 órányi szünetet tartalmaz, vagyis a teljes korpusz 10,9%-át a szünetek teszik ki. Az általunk létrehozott VAD algoritmust 5 órányi tanító adatbázison készítettük el, a tanító adatbázis a küszöb beállítására szolgált. A VAD kialakítása után 36 órányi spontán beszéden futtattuk az algoritmust a teszteléséhez.

Az első kísérletben azt teszteltük, hogy a milyen hosszú ablakhosszt kell optimálisan választani ahhoz, hogy a legjobb felismerési eredményt kapjuk. Az ablakhosszt 1-től 5 keretig növeltük, vagyis 25 ms-tól 250 ms-ig. Ezzel egy időben azt is teszteltük, hogy melyik módszerrel (az alap vagy az általunk javasolt) tudunk elérni jobb detektálási eredményt. A legkisebb hibát akkor kaptuk, hogy ha a szegmentáláshoz 5 keret, vagyis 250 ms-os hosszúságú ablakot használtunk. Ekkor a szegmentálási hiba értékéke 9,51%. Mindemellett az eredményekből az is látszik, hogy az általunk javasolt k-középpel működő szegmentáló 3 keret hosszúságú ablaktól jobb eredményt ad, mint az alaprendszer, azonban ez a különbség nem szignifikáns.

A tesztelés során megvizsgáltuk azt is, hogy az általunk javasolt VAD milyen eredménnyel működik különböző jel/zaj viszonyú (SNR: Signal to Noise Ratio) spontán beszédben. Rendszerünk zajtűrését úgy teszteltük, hogy a felvételhez fehér zajt kevertünk a jel/zaj arányt folyamatosan csökkentve ezzel. A felvétel eredeti SNR értéke átlagosan 25%. Az SNR értékét 5 dB-enként csökkenetettük. A zaj hatására az eredményeink csökkennek, azonban még 10% SNR mellet is 34,72%-os a hiba értéke.

A jelen vizsgálat eredményei alapján kimondható, hogy az általunk javasolt módszerrel a beszéddetektálási hiba csökkenthető, statisztikailag azonban a javulás nem volt igazolható. Az általunk javasolt rendszer jó minőségű felvételen 90,49%-os eredménnyel működik. 10%-os jel/zaj arányig még közel 65,28%-os eredménnyel, 5%-os jel/zaj aránytól viszont már csak 38,8%-os helyes találati aránnyal működik a rendszer.

4.2 EGYSZERRE BESZÉLÉSEK AUTOMATIKUS OSZTÁLYOZÁSA

Az egyszerre beszélések aránya a spontán társalgásokban meglehetősen nagynak mondható (Gráczi–Bata 2010). Beattie a beszélőváltásokat elemezve (1983, idézi Levelt 1989) kimutatta, hogy a két résztvevős angol társalgásban 11%-ban fordul elő egyszerre beszélés, több beszélőnél ez az arány már 31%. Cetin és Shriberg (2006) angol korpuszokat vizsgálva azt adatolta, hogy az átfedő beszéd átlagosan 10-13%-át teszi ki a társalgásoknak. A hazai kutatásokban Markó (2006) 6%-ot állapít meg a teljes beszéd és az átfedő beszéd arányaként négybeszélős spontán társalgásban, Bata (2009) pedig 1,7-3%-ot adatolt spontán társalgásokban.

A beszélődetektálásban kimutatták, hogy a legtöbb hiba szignifikánsan azon részeken történik a felvételekben, ahol egyszerre beszélés található. Wooters és Huijbert (2007) szerint a beszélődetektálási hiba arányának (DER) 17%-át a téves elutasítások száma adja, amelyet az átfedő beszédrészek okoznak. Az egyszerre beszéléseket modellező munkák száma relatíve kevés, és azok közül is csak néhány kutatásban mutatták ki, hogy az egyszerre beszélés detektálása csökkenti a beszélődetektálási hiba arányát (Boakye et al. 2008 a,b,c; Boakye 2008; Boakye 2011; Trueba-Hornero 2008).

A jelen kutatás célja, hogy a spontán társalgásokban modellezze az egyszerre beszéléseket, és automatikus osztályozó algoritmussal különítse el azoktól a beszédszakaszoktól, ahol csak egy társalgó beszél. Hipotézisünk szerint az átfedő beszéd jellegzetes akusztikai szerkezettel rendelkezik, ezért létrehozható egy automatikus osztályozó algoritmus. Ugyanakkor feltételezzük, hogy a háttércsatorna-jelzések okozzák majd a legtöbb hibát az osztályozáskor.

A jelen kutatásban egy ANN/SVM hibrid rendszer (Mesterséges Neuron Hálók/Szupport Vektor Gépek, Artificiel Neural Network/Support Vector Machine) hoztunk létre az egyszerre beszélések automatikus osztályozásához.

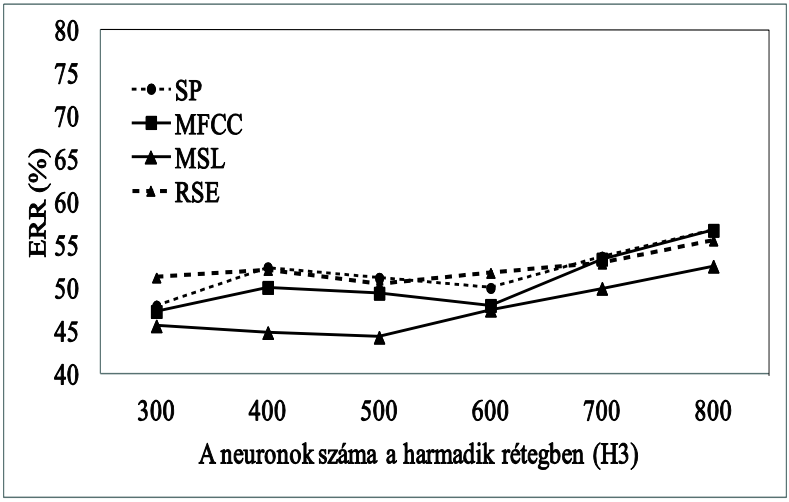
Az osztályozás első lépése a lényegkiemelés, amely során a beszédjelből olyan információkat vonunk ki, amellyel jól megragadhatók az egyszerre beszélések. Négy akusztikai jellemzőt teszteltünk: spektrum, mel-frekvencia kepsztrális (MFC) együtthatók, Mel-skála szerinti logaritmikus filterbank (MSL), részsáv-energiát (RSE). A jellemzők kinyerése után korlátozott Boltzmann-géppel (restricted Boltzman machine, RBM) emeltük ki a lényeget az akusztikai jellemzőkből (Dahl et al. 2010), majd a rejtett rétegek aktivációs értékeit használtuk fel az átfedő beszédrészek és nem-átfedő beszédrészek automatikus osztályozásához, amelyet Szupport Vektor Géppel (RBF kernelfüggvénnyel) valósítottunk meg. Mélyrétegű neurális hálózatok létrehozásához 1-3 RBM-et (H1, H2, H3) kapcsoltunk össze úgy, hogy a megelőző rejtetett réteg aktivációja a következő látható réteg bemenete.

A társalgásokban manuálisan jelöltük azokat a részeket, ahol egyszerre több adatközlő beszél, illetve azokat a részeket, ahol csak egy beszélő beszél. A 100 beszélő spontán társalgásaiban összesen 8056 olyan időintervallum található, ahol kettő vagy annál több résztvevő szólal meg egyszerre. Az átfedő beszéd összidőtartama közel 7 óra, amely a teljes korpusz 12%-a. Az SVM tanításához a 8056 átfedő beszédszegmens 2/3-át, vagyis 5370-et használtunk fel, míg a teszteléshez az 1/3-át, amely 2386-ot. A bemenő jellemzővektorokat azonos dimenziójúra hoztuk úgy, hogy az egyes audioszegmensek kereteire statisztikai jellemzőket számolunk (átlag és szórás).

A jelen kutatásban teszteltük, hogy a négy akusztikai paraméter közül melyikkel lehet elérni a legjobb eredményt. Továbbá teszteltük azt is, hogy hogyan változik az eredményünk annak függvényében, hogy a mélyrétegű neuronhálózat harmadik rétegében hány neuront használunk.

Az eredmények azt mutatják, hogy a négy akusztikai paraméter közül a legjobb teljesítményt akkor kaptuk, ha jellemzőként a mel-skála szerinti logaritmikus filterbank-ot alkalmaztuk. Ekkor az Equal Error Rate (EER) átlagos értéke 47,49%, vagyis az helyesen felismert szegmensek száma átlagosan 52,51%. A második legjobban teljesítő jellemző az MFCC volt. Ennek átlagos EER értéke 50,84% volt. Elmondható tehát az, hogy átlagosan 3,35%-os hiba csökkenést tudtunk elérni a MSL jellemző alkalmazásával az MFCC-vel elért eredményhez képest. Ez a javulás szignifikáns (Wilcoxon próba: Z=-2,211; p=0,023).

Megvizsgáltuk, hogy az EER értéke hogyan függ a jellemzők és a harmadik rétegben használt neuronok számától. Az eredmények azt mutatják, hogy a legjobb eredményt akkor kapjuk, ha MSL jellemzőt és 500 neuront használunk a H3-ban (1. ábra).



1. ábra.
Az EER értéke a jellemzők és a H3-ban lévő neuronok számának függvényében

A statisztikai elemzések alátámasztják, hogy a MSL szignifikánsan jobban teljesít attól függetlenül, hogy hány neuront használunk a harmadik rétegben: MSL-MFCC: Z=-2,201; p=0,028; MSL-SP: Z=-2,201; p=0,028; MSL-RSE: Z=-2,201; p=0,028.

Az EER értékekből azt látszik, hogy két esetben (spektrum és MFCC) akkor volt a legkisebb a hiba értéke, ha a harmadik rétegben 300 neuront használtunk. Az MSL és a RSE esetében pedig a legkisebb hibát akkor kaptuk, ha a neuronok száma 500 volt a harmadik rétegben. Általánosságban azonban az elmondható, hogy 500 neuron felett mindegyik jellemző esetében nőtt az EER értéke.

Az elért eredményeinket visszaellenőrizve elemeztük a hibák tulajdonságait. Az első és legnagyobb hibaforrás maga a kézi címkézés volt. Emellet sok hibát okoztak az igen rövid időtartamú háttércsatorna-jelzések és a nevetések.

4.3 BESZÉLŐFELISMERÉS A BESZÉLŐDETEKTÁLÁSHOZ

Az utóbbi évtizedekben egyre nagyobb figyelmet kapó automatikus beszélőfelismerő rendszerek számos más beszédtechnológiai alkalmazásba integrálhatók, ilyen például a beszédfelismerés, de a napjainkban a legdinamikusabban fejlődő beszélődetektálónak (speaker diarisation) is szerves része (Campbell 1997). A magyar nyelvre vonatkozóan számos kutatás jelent meg a beszélő személy azonosításának témakörében (Gósy–Nikléczy 1999; Nikléczy 2003; Beke 2008; Bőhm 2006). Igen kevés számban jelent meg azonban kifejezetten a beszélő személy gépi felismerésével foglalkozó tanulmány (Fék 1997).

A jelen kutatás célja kettős: (i) megvizsgálni, hogy a magyar nyelvű beszédben mely spektrális régiók beszélőspecifikusak, (ii) a beszélőket MFC-vel előfeldolgozva GMM-ekkel, illetve GMM-UBM-ekkel modellezni és osztályozni a spontán beszédük alapján. A szövegfüggetlen beszélőosztályozó eredményeit az általunk fejlesztett beszélődetektálóba kívánjuk integrálni.

Az egyes beszélők automatikus felismeréséhez különböző számú Gauss-komponenst tartalmazó GMM-eket használtunk a tanítás és a tesztelés során. A tanításhoz 80 beszélő 25 s-os beszédmintáit használtuk. A tesztelést 13 s-os beszédmintán végeztük el. A tanítás során minden egyes beszélőre külön modellt hoztunk létre. Az általános háttérmodell (UBM) kialakításához a tanító adatbázistól a további 20 adatközlő 25 s-os beszédét használtunk fel. A kutatás során teszteltük, hogy mely MFC együtthatóval a legsikeresebb az osztályozás: i) teljes spektrumot kódoló MFC; ii) 1,5–2,5 kHz közötti MFC, iii) 2,5–3,5 kHz közötti MFC; iv) 3,5–4,5 kHz közötti MFC.

A beszélők modellezéséhez kevert Gauss-modelleket és a beszélő valószerűségi értékét normalizáló általános háttérmodellt használtunk (Higgins et al. 1991; Rosenberg et al. 1992; Reynolds 1995; Matsui–Furui 1995; Reynolds 1997). A beszélőazonosításra a likelihood ratio test-et (valószerűségi arány teszt) szokás alkalmazni az azonosítandó beszédekre. A beszélő azonosításakor a háttérmodell egy részét a beszélőmodellekből kell előállítani. Ebben a vizsgálatban a 20 beszélőből készítettük el a háttérmodellt. Azt a valószínűséget, ami nem az egyes beszélőktől származik, hanem a tanító adatbázisban szereplő beszélőktől, általános háttérmodellnek hívjuk (universal background model). A beszélőfelismerő rendszer kiértékelésére a felismerés pontosságát (Accuracy), vagyis a helyesen felismert adatok arányát adtuk meg.

Az eredmények egyfelől azt mutatták, hogy a Gauss-komponensek függvényében a valószínűségi érték az azonos beszélők esetében egyre magasabb értéket vesz fel, míg a különböző beszélők esetében ez az érték csökken.

Másfelől, a GMM-et általános háttérmodell használatával átlagosan jobb eredményeket kaptunk, mint a GMM általános háttérmodell nélkül. A statisztikai elemzés szerint ez a különbség szignifikáns: Wilcoxon próba: Z=-2,944; p=0,003.

Megvizsgáltuk azt is, hogy mely akusztikai jellemzővel használt osztályozó adja a legjobb eredmény. A legjobb osztályozási arányt a 2500-3500 Hz részsávra számolt MFC együtthatókkal értük el mind a GMM, mind a GMM-UBM esetében. Ez azonban statisztikailag csak részben igazolható. Az $MFC_{(2,5–3,5)}$ jellemzővel elért eredmények szignifikánsan különböznek az $MFC_{(1,5–2,5)}$-vel (Z=-2,201; p=0,028) és az $MFC_{(3,5–4,5)}$-vel (Z=-2,201; p=0,028) elért eredményektől, azonban a teljes spektrumot lekódoló eljárástól nem. Az adatokból azonban látszik, hogy szisztematikusan jobban teljesít az $MFC_{(2,5–3,5)}$ jellemző, mint az $MFCC_{(full-ban)}$. Ez az eredmény megerősíti a nemzetközi kutatások eredményeit, miszerint valóban a spektrum ezen régiója (2,5 kHz és 3,5 kHz) hordozza az egyéi beszédjellemzőket.

Elemeztük továbbá, hogy a felismerés pontossága hogyan alakul a Gauss-komponensek függvényében. Az eredményekből az látszik, hogy a Gauss-komponensek számának növekedésével javul a pontosság értéke is.

Összességében tehát elmondható, hogy a legjobb eredményt az $MFC_{(2,5-3,5)}$ jellemzőt használó 256 Gauss-komponenst tartalmazó GMM-UBM osztályozóval értük el, amelynek értéke 79,76%-volt. Az eredményeink azt is mutatják, hogy a Nikléczy–Gósy (2008) által megállapított 16-s-nál rövidebb, 13 s-os rész is elégséges ahhoz, hogy a beszélőket alacsony hiba aránnyal tudjuk automatikusan felismerni a beszédhang alapján.

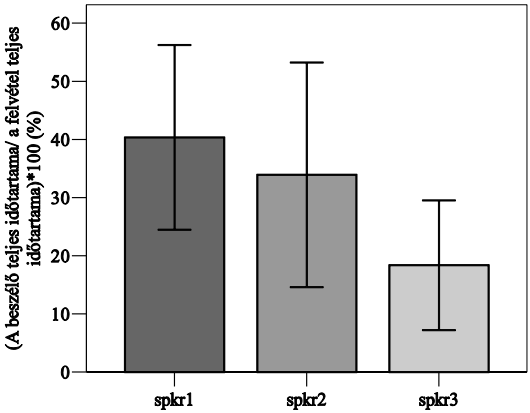
4.4 AUTOMATIKUS BESZÉLŐDETEKTÁLÁS

Az eddigi beszélődetektáló rendszereket rádiós műsorokon hozták létre különböző nyelveken, amelyek nagyrészt félspontánnak minősülnek, hiszen a műsor résztvevői előzetesen ismerik a témát. A BEA adatbázis spontán társalgásaiban a beszédtervezés és kivitelezés egyszerre zajlik ott helyben, ezért azt mondhatjuk, hogy a jelenleg használt korpusz jobban közelít a spontán beszédhez, mint az eddig használt korpuszok. Ezért a jelen dolgozat szintén újszerűnek mondható, mivel ilyen jellegű spontán társalgásokon való beszélődetektáló kialakítása eddig még nem történt meg.

A BEA adatbázisból általunk random kiválasztott 100 társalgásban összesen 7827 db beszédforduló volt. Egy felvételre átlagosan 78 db beszédforduló jut, amelynek szórása 41 db. Egy felvételen belül a legtöbb beszédforduló 240 db volt, míg a legkevesebb 11 db. Megvizsgáltuk, hogy a nemek között van-e különbség a beszédfordulók gyakoriságának tekintetében. Azokban a társalgásokban, amelyekben férfi volt az adatközlő, átlagosan 79 db (szórás 45 db) beszédforduló volt, míg ahol nő, 65 db (szórás 37 db), ez a különbség azonban nem szignifikáns (egytényezős ANOVA).

Megvizsgáltuk, hogy a társalgásokon belül az egyes beszélők a teljes időtartamra nézve hány százalékában szólalnak meg. Az adatok szerint az adatközlők átlagosan 40,3%-ban tartják maguknál a szót. A férfi adatközlők átlagosan 37%-ban tartják

magunknál a szót a teljes időtartamhoz képest, míg a nők 42%-ban, azonban ez a különbség szintén nem szignifikáns (egytényezős ANOVA). A felvételvezető átlagosan 33,9%-ban tartja magánál a szót, míg a harmadik résztvevő csupán átlagosan 18,3%-ban (2. ábra). Ezek az arányok azt mutatják, hogy a társalgások során a szerepek nem kiegyenlítettek, a harmadik személy sokszor háttérbe szorul (ennek oka többféle lehet, pl.: ismertségi fok). Ez a kiegyenlítetlenség statisztikai elemzéssekkel is alátámasztható, hiszen mind a felvételvezető, mind az adatközlő szignifikánsan többet beszél, mint a harmadik résztvevő (ismétléses ANOVA: spkr1*spkr3: $F(2, 200)=39,833$; $p<0,001$; spkr2*spkr3 $F(2, 200)=39,833$; $p<0,001$).



2. ábra.
A beszélők teljes időtartama a felvétel teljes időtartamának függvényében

Kiszámoltuk, hogy az egyes résztvevőkre hány beszédforduló jut egy percre. Az adatközlőre átlagosan 1,38 beszédforduló jut egy percre, a felvételvezetőre 1,15, míg a harmadik esetében 0,78. Ez szintén a társalgás résztvevőinek aszimmetriáját mutatja.

Továbbá megvizsgáltuk, hogy a beszédidőtartamok és a beszédforduló/perc hogyan függnek össze az egyes résztvevők függvényében. Az adatközlőknél nem lehet kimutatni semmilyen tendenciát, vagyis e két jelenség nem függ össze egymással; tehát nem lehet azt mondani, hogy aki sokat beszél, az többször kap, vagy veszi át a szót. A kísérletvezető esetében azonban pozitív közepesen erős függvénykapcsolatot tudtunk kimutatni (Pearson korreláció: r =0,424, p<0,001). Ugyanilyen tendenciát találtunk a harmadik résztvevő esetében is (Pearson korreláció: r =0,441, p<0,001). Mindez azt mutatja, hogy míg az adatközlőknek nem kell törekedni a szóátvételre, hiszen az alaphelyzet az, hogy ő beszéljen, addig a felvételvezetőnek és a harmadik személynek igen, vagyis ahhoz, hogy minél hosszabb közléseket hozzanak létre, annál többször kell magukhoz venniük a szót.

Az általunk javasolt beszélőszegmentáláshoz a *Beszélőfelismerés a beszélődetektáláshoz* fejezetben bemutatott MFCC eljárást használtuk jellemzőkinyerő algoritmusként (először a teljes spektrumra, majd 2,5 kHz és a 3,5 kHz közöttire). A beszélőklaszterezés során a beszélőszegmentálóból érkező szegmensek a bemenet, vagyis a két beszélőváltás között lévő beszédjel. Ezen beszédjelek modellezésére a kinyert akusztikai jellemzőkből GMM-szupervektorokat képeztünk (MAP adaptált középértékek összefűzésével, vö. Reynolds et a. 2000). A jelen kutatásban a GMM-UBM 256 kevert komponenst tartalmaz. Mivel így egy magas-dimenziószámú jellemzővektort kapunk, ezért a jelen dolgozatban a dimenziószámot lecsökkentettük PCA-val. A dimenziócsökkentett jellemzővektor jelen esetben az i-vektror. Az i-vektor a bemenet a nem-ellenőrzött tanuláson alapuló beszélőklaszterezésnek.

A jelen munkában a beszélők szegmentáláshoz a Bayesian Information Criterion (BIC) algoritmust használtuk, amely a feltételes valószínűségszámítás alapjain nyugszik. A BIC-ben a modellkiválasztás úgy történik, hogy a valószínűségi kritérium érték annál magasabb, minél magasabb a modell komplexitása, tehát bünteti a modell komplexitást (szabad paraméterek összege a modellben) (Schwarz 1971; 1978). A BIC algoritmust elsőként Chen és Gopalakrishnan (1998) alkalmazta a beszélődetektálásban, ahol egy teljes kovarianciájú Gausst használtak az adatok modellezéséhez. Bár nem létezik eredeti formula, a λ paraméter úgy van bevezetve, mint a büntetőfaktor hatása az összehasonlításban, amely rejtett küszöbértéket alkot a BIC különbséghez. Mivel a küszöbérték megválasztása fontos az adatok illesztéséhez, ezért számos tanulmány foglalkozott azzal, hogy milyen módszerrel lehet ezt a szabad paramétert optimálisan megválasztani. Néhány tanulmány amellett érvel, hogy automatikusan kell a λ paramétert megválasztani (Tritschler–Gopinath 1999; Delacourt–Wellekens 2000; Delacourt–Kryze–Wellekens 1999; Mori–Nakagawa 2001; Lopez–Ellis 2000; Vandecatseye et al. 2004).

A BEA adatbázisból 12 társalgást választottunk ki random módszerrel. A 12 társalgás összidőtartama közel 2,8 óra, amelyben 480 beszédváltás történt. A beszélődetektáló kiértékelésénél minden tesztfájlra meghatároztuk a DER értéket (*diarization error rate*).

A standard BIC beszélődetektáló rendszerben MFCC teljes spektrumot lekódoló jellemzőt használunk, a BIC λ paraméterét 0-ra állítottuk, és nem használtunk sem szünetmodellt, sem egyszerre beszélés modellt a beszélődetektáláshoz. Ennek átlagos eredménye 39,43%-os DER, ami azt jelenti, hogy a 60,56%-ban helyesen szegmentál és klaszterez az alap kiinduló algoritmusunk.

A továbbiakban az általunk javasolt beszélődetektálóba integráljuk az egyes fejezetekben bemutatott blokkokat, illetve teszteljük annak hatását. Célunk annak bemutatása, hogy a különálló blokkokban kikísérletezett eredmények hogyan hasznosíthatók a beszélődetektálásban.

A *Beszélőfelismerés a beszélődetektáláshoz* című fejezetben bemutattuk, hogy ha az MFCC jellemzőkinyerést 2,5 kHz és 3,5 kHz-es részsávban végezzük, akkor a beszélőszemély-felismerés eredménye növelhető, hiszen ez a frekvenciatartomány

tartalmazhatja a beszélőre specifikus akusztikai lenyomatokat. Ezt az akusztikai paramétert teszteltük a beszélődetektálóban is. A beszélődetektálóban elért eredmények szintén igazolták, hogy az $MFCC_{(2,5–3,5)}$ akusztikai jellemző átlagosan jobban teljesít, mint az MFCC a teljes spektrumra számolva. A $MFCC_{(2,5–3,5)}$ jellemzővel 38,56% DER értéket kaptunk, amely átlagosan 0,869%-so DER javulást okozott. Jóllehet az átlagos javulás mértéke alacsony, ez a különbség szignifikáns Wilcoxon teszt szerint (Monte Carlo szimulációval kiegészítve: Z=-2,824; p=0,005).

Teoretikusan a λ büntető faktor értéke zéró, amely a gyakorlatban sokszor 1-re szokás állítani (Ajmera et al. 2004). A jelen dolgozatban 0-tól 4-ig növeltük a λ paraméter értékét és megvizsgáltuk, hogy hogyan változok a DER értéke. Az akusztikai jellemzőként az $MFCC_{(2,5–3,5)}$-t használtuk. A tesztelés során a legjobb eredményt, vagyis a legkisebb DER értéket akkor kaptuk, ha a BIC λ paraméterét 1-re állítottuk. Ekkor az átlagos beszélődetektálás hiba aránya 35,73%.

A *Beszéddetektálás* című fejezetben létrehozott VAD-ot implementáltuk a beszélődetektálóba. Az eljárás lényege, hogy a VAD által detektált szünet részeket már nem továbbítottuk a beszélődetektáló felé, vagyis töröltük felvételből. Tehát jelen esetben a VAD-ot mint előfeldolgozó egységként csatoltuk a beszélődetektáló elé. Az eredmények azt mutatják, hogy a VAD előfeldolgozásával az DER értékét átlagosan 4,535%-al tudtuk csökkenteni. Ez az átlagos javulás statisztikailag igazolható (Wilcoxon teszt, Monte Carlo szimulációval kiegészítve: Z=-3,059; p<0,001).

Az *Egyszerre beszélések automatikus osztályozása spontán magyar társalgásokban* című fejezetben létrehozott algoritmust implementáltuk a beszélődetektáló rendszerünkbe. Hasonlóan a VAD-hoz az egyszerre beszélés detektálót úgy alkalmaztuk, hogy az általa generált kimenet alapján a társalgásból kivágtuk azon részeket, ahol egyszerre több beszélő szólalt meg. Tehát jelen esetben az egyszerre beszélés detektálót mint előfeldolgozó egységként csatoltuk a beszélődetektáló elé a VAD egység után. Az átfedő beszédek automatikus detektációjával átlagosan 2,49%-os relatív javulást tudtunk elérni, vagyis a DER értékét 31,21%-ról le tudtuk csökkenteni 28,71%-ra. Ez a javulás szignifikáns (Wilcoxon teszt, Monte Carlo szimulációval kiegészítve: Z=-3,06; p=0,002).

Összességében elmondható, hogy a legjobb eredményt akkor kaptuk, ha a BIC λ paraméterét 1-re állítottuk, $MFCC_{(2,5–3,5)}$ akusztikai jellemzőt alkalmaztunk és előfeldolgozásként implementáltuk mind a VAD, mind az egyszerre beszélés detektáló algoritmusokat. Ekkor a DER értéke 28,71% volt.

5. KÖVETKEZTETÉSEK

A jelen kutatás fő célja az volt, hogy magyar nyelvre elsőként hozzon létre spontán társalgásokra nem-ellenőrzött tanuláson alapuló beszélődetektáló algoritmust. A kutatás egyik fő kérdése az volt, hogy milyen eredménnyel tudjuk megvalósítani a beszélődetektálót a spontán társalgásokra. Hogyan valósíthatók meg a beszélődetektálás egyes előfeldolgozó rendszerei, mint a beszéddetektálás, egyszerre

beszélés detektálás, illetve hogy ezek milyen eredménnyel implementálhatók a beszélődetektáló rendszerbe. Arra is kerestük a választ, hogy melyek azok az akusztikai jellemzők, amelyek az egyénre jellemző akusztikai lenyomatokat tartalmazhatják. Vizsgáltuk, hogy milyen eredménnyel lehet a képi feldolgozásban használt mély neuronhálókat alkalmazni az egyszerre beszélés detektáló jellemzőkinyeréseként. Elemeztük, hogy a beszélőszegmentálásban milyen beállítások mellet kapjuk a legjobb eredmény.

1. Az általunk javasolt beszéddetektáló rendszer jó minőségű felvételen 90,49%-os eredménnyel működik; 10%-os jel/zaj arányig még közel 65,28%-os eredménnyel, 5%-os jel/zaj aránytól viszont már csak 38,8%-os helyes találati aránnyal. Ez azzal magyarázható, hogy a VAD-algoritmusban nem használtunk zajszűrőt. Ezért tervezzük, hogy zajszűrőkkel is kísérletezni fogunk. Az elkészített VAD egy általunk fejlesztett beszélődetektálóba lesz integrálva, amely feltehetőleg javítani fogja annak működését.

2. Az egyszerre beszélések detektálása során a legjobb eredményt a Mel-skála szerinti logaritmikus filterbank jellemző adta. Ez korrelál más kutatásokban is ezt a jellemzőt használó algoritmusok által elért eredménnyel, például beszédhang-felismerésben (Li et al. 2012; Mohamed et al. 2012). Ezen tanulmányok arról számoltak be, hogy a Mel-skála szerinti logaritmikus filterbank jellemző jobban teljesített, mint az MFCC. A legjobb eredményt akkor kaptuk, az EER értéke 44,33%, ha Mel-skála szerinti logaritmikus filterbank jellemzőt, és H1(300)-H2(600)-H3(500) topológiájú DBN-t használtunk előfeldolgozásként, és SVM-RBF-et osztályozóként.

3. A beszélő személy felismerésben az eredmények azt mutatják, a spektrumban a 2,5 kHz és a 3,5 kHz közé eső frekvencia tartomány őrzi a beszélő személyre utaló akusztikai jegyeket. Ez az eredmény megerősíti a nemzetközi kutatások eredményeit, miszerint valóban a spektrum ezen régiója hordozza az egyéni beszédjellemzőket. A legjobb eredményt akkor értük el, ha 256 komponenst tartalmazó GMM-UBM-et használtunk, amelynek értéke 79,76%-volt.

4. Bemutattuk, hogy ha az MFCC jellemzőkinyerést 2,5 és 3,5 kHz-es részsávban végezzük, akkor a beszélőszemély-felismerés eredménye növelhető. Ezt az akusztikai paramétert teszteltük a beszélődetektálóban is. A beszélődetektálóban elért eredmények szintén igazolták, hogy az $MFCC_{(2,5–3,5)}$ akusztikai jellemző átlagosan jobban teljesít, mint az MFCC. A $MFCC_{(2,5–3,5)}$ jellemzővel 38,56% DER értéket kaptunk, amely átlagosan 0,869%-so DER javulást okozott. A tesztelés során a legjobb eredményt, vagyis a legkisebb DER értéket akkor kaptuk, ha a BIC λ paraméterét 1-re állítottuk. Ekkor az átlagos beszélődetektálás hiba aránya 35,73%

4.1 Az eredmények azt mutatták, hogy a VAD előfeldolgozásával az DER értékét átlagosan 4,535%-al csökkenthető.

4.2 Az átfedő beszédek automatikus detektációjával átlagosan 2,49%-os relatív javulást tudtunk elérni, vagyis a DER értékét 31,21%-ról le tudtuk csökkenteni 28,71%-ra.

5. Összességében elmondható, hogy a legjobb eredményt akkor kaptuk, ha a BIC λ paraméterét 1-re állítottuk, $MFCC_{(2,5–3,5)}$ akusztikai jellemzőt alkalmaztunk és előfeldolgozásként implementáltuk mind a VAD, mind az egyszerre beszélés detektáló algoritmusokat. Ekkor a DER értéke 28,71% volt.

A beszélődetektáló hasznos lehet mind a nyelvészek, mind a beszédtechnológusok számára. A nyelvészek használhatják a konverzációelemzéshez, hiszen automatikusan lehet a rendszerrel a társalgásokat beszélők szerint annotálni. A beszélődetektálás a beszédtechnológiában, azon belül a beszédfelismerésben a beszélőadaptált rendszerek megalkotásában játszhat fontos szerepet. A disszertáció eredményei közelebb vihetik a kutatót az ember-ember kommunikáció megértéséhez, modellezéséhez, amely tovább mutat a mesterséges intelligencia, az ember-gép kommunikációja felé.

6. ÖSSZEOGLALÁSHOZ FELHASZNÁLT IRODALOM

Bata Sarolta 2009. Beszélőváltások a beszédpartnerek személyes kapcsolatának függvényében. In: Beszédkutatás 2009. 107–120.

Beattie, G. W. 1982. Turn-taking and interruption in political interviews: Margaret Thatcher and Jim Callaghan compared and contrasted. Semiotica 39: 93–114.

Beke András 2008. Az alapfrekvencia-eloszlás modellezése a beszélőfelismeréshez. Alkalmazott Nyelvtudomány 2008/1–2: 121–132.

Boakye K. 2008. Audio Segmentation for Meetings Speech Processing. Ph.D. dissertation, University of California at Berkeley, 2008.

Boakye K. – Vinyals O. – Friedland G. 2008a. Two'sa crowd: Improving speaker diarization by automatically identifying and excluding overlapped speech. In Proc. Interspeech 2008. 32–35.

Boakye K. –Trueba-Hornero B. –Vinyals O. –Friedland G. 2008b. Overlapped speech detection for improved speaker diarization in multiparty meetings. Proc. ICASSP. 4353–4356, 2008.

Boakye, K. – Trueba-Hornero, B. – Vinyals, O. – Friedland, G. 2008c. Overlapped speech detection for improved speaker diarization in multiparty meetings. In: Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing. Las Vegas, Nevada. 4353–4356.

Boakye, K. – Vinyals, O. – Friedland, G. 2011. Improved Overlapped Speech Handling for Speaker Diarization. In: Proceeding of INTERSPEECH 2011. Firenze, Olaszország. 941–944.

Bőhm Tamás 2006. A glottalizáció szerepe a beszélő személy felismerésében. Beszédkutatás 2006. 197–207.

Campbell, J. P. 1997. Speaker Recognition: A Tutorial. In: *Proceedings of the the Institute of Electrical and Electronic Engineers*, Vol. 85, No. 9. 1437–1462.

Çetin, Ö. – Shriberg, E. 2006. Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site: insights for automatic speech recognition. In: *Proceedings of INTERSPEECH 2006*. 293–296.

Cettolo, M. – Vescovi, M. 2003. Efficient audio segmentation algorithms based on the BIC. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*.

Chen, S. S. – Gopalakrishnan, P. 1998. Clustering via the Bayesian information criterion with applications in speech recognition. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing.* Seattle, USA, 645–648.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, Letöltés ideje: 2013.06.05. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Cho, Y.D.; Kondoz, A. (2001). Analysis and improvement of a statistical model-based voice activity detector, IEEE Signal Processing Letters, vol. 8, no. 10, 276–278.

Dahl, G. E., Ranzato, M., Mohamed, A., Hinton, G.: Phone recognition with the mean-covariance restricted boltzmann machine. In: NIPS (2010) 469–477.

Daniel P. W. Ellis 2005. PLP and RASTA and MFCC, and inversion in Matlab, http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/

Delacourt, P. – Kryze, D. – Wellekens, C. J. 1999. Detection of speaker changes in an audio document. In: *Proceedings of Eurospeech 1999*. 1195–1198.

Delacourt, P. – Wellekens, C. J. 2000, DISTBIC: A speaker-based segmentation for audio data indexing. *Speech Communication: Special Issue in Accessing Information in Spoken Audio* 32: 111–126.

Fék Márk 1997. Beszélőfelismerés neurális hálózatokkal és vektorkvantálóval. *OTDK konferencia*. Szeged 1997.

Garfinkel, H. 1967. *Studies in Ethnomethodology*. Prentice Hall, Englewood Cliffs, NJ.

Giannakopoulos, T. 2009. Study and application of acoustic information for the detection of harmful content, and fusion with visual information. Department of Informatics and Telecommunications,University of Athens, Greece, PhD theis.

Goffman, E. 1983. The Interaction Order. *American Sociological Review* 48: 1–17.

Gósy Mária 2012. Multifunkcionális beszélt nyelvi adatbázis – BEA. In Prószéky Gábor – Váradi Tamás (szerk.): Általános Nyelvészeti Tanulmányok XXIV. Nyelvtechnológiai kutatások. Akadémiai Kiadó, Budapest, 329–349.

Gráczi Tekla Etelka – Bata Sarolta 2010. Megszólalási formák és funkciók az összeszokottság függvényében. In: Gecső Tamás – Sárdi Csilla (szerk.) *Új módszerek az alkalmazott nyelvészeti kutatásban.* Kodolányi János Főiskola, Tinta Könyvkiadó, Székesfehérvár, Budapest. 28–32.

Higgins, A. L. – Bahler, L. – Porter, J. 1991. Speaker verification using randomized phrase prompting. *Digital Signal Processing* 1/2: 89–106.

Hung, J. – Wang, H. – Lee, L. 2000. Automatic metric based speech segmentation for broadcast news via principal component analysis. In: *Proceedings of the International Conference on Speech and Language Processing*, Beijing, China.

Ida, O. 2011. Indexing of Audio Databases : Event Log of Broadcast News. PhD thesis. Norwegian University of Science and Technology, Department of Electronics and Telecommunications.

Iványi Zsuzsanna 2001. A nyelvészeti konverzációelemzés. *Magyar Nyelvőr 125*. 74–93.

Jin, Q. – Laskowski, K. – Schultz, T. – Waibel, A. 2004. Speaker segmentation and clustering in meetings. In: *Proceedings of NIST 2004 Spring Rich Transcrition Evaluation Workshop*, Montreal, Canada. 112–117.

Markó Alexandra 2006. Beszélőváltások a társalgásban. http://fonetika.nytud.hu/letolt/ma_2.pdf (Letöltve: 2011. október 1.)

Matsui, T. – Furui, S. 1995. Likelihood normalization for speaker verification using a phoneme- and speaker-independent model. *Speech Communication* 17: 109–116.

Matza, A.– Bistritz, Y. 2011. Skew Gaussian mixture models for speaker recognition. Presentation. In: *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH) 2011*. 28–31.

Mohamed, G. Hinton, and G. Penn, 2012. Understanding how deep belief networks perform acoustic modelling. In: Proc. ICASSP, 4273-4276, 2012.

Mori, K. – Nakagawa, S. 2001. Speaker change detection and speaker clustering using VQ distortion for broadcast news speech recognition. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Salt Lake City, USA, 413–416.

Németh Géza – Olaszy Gábor (szerk.) 2010. A magyar beszéd. Beszédkutatás, beszédtechnológia, beszédinformációs rendszerek (8–12. fejezet). Akadémiai Kiadó, Budapest.

Nikléczy Péter – Gósy Mária 2008. A személyazonosítás lehetősége a beszédanyag időtartamának függvényében. *Beszédkutatás 2008*. 172–181.

Nikléczy Péter 2003. A zönge periódusidejének funkciója a hangszínezetben. *Beszédkutatás 2003*. 101–113.

Reynolds, D. A. – Quatieri, T. F. – Dunn, R. 2000. Speaker verifcation using adapted Gaussian mixture models. *Digital Signal Processing* 10/1–3: 19–41.

Reynolds, D. A. 1995. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication* 17: 91–108.

Reynolds, D. A. 1997. Comparison of background normalization methods for text-independent speaker verification, In: *Proceedings of 5th European Conference on Speech Communication and Technology (Eurospeech).* 963–966.

Rosenberg, A. E. – DeLong, J. – Lee, C.-H. – Juang, B.-H. –Soong, F. K. 1992. The use of cohort normalized scores for speaker verification. In: *Proceedings of International Conference on Spoken Language Processing*. 599–602.

Sacks, H. – Schegloff, E. A. – Jefferson, G. 1974. A simplest systematics for the organization of turntaking for conversation. *Language* 50: 696–735.

Sacks, H. 1992. *Lectures on Conversation*. Blackwell, Oxford.

Schegloff, E. 1992. Introduction. In: Sacks, H. *Lectures on Conversation*. Vol.1. Blackwell, Oxford. 9–12.

Schwarz, G. 1971. A sequential student test. *The Annals of Statistics* 42/3: 1003–1009.

Shih-Sian Cheng – Hsin-Min Wang – Hsin-Chia Fu 2010. BIC-Based Speaker Segmentation Using Divide-and-Conquer Strategies With Application to Speaker Diarization," IEEE Transaction on Audio, Speech, and Language Processing, vol. 18, no. 1, 141–157, Jan. 2010.

Siegler, M. A. – Jain, U. – Raj, B. – Stern, R. M. 1997. Automatic segmentation, classification and clustering of broadcast news audio. In: *Proceedings of DARPA Speech Recognition Workshop*, 97–99.

Sivakumaran, P. – Fortuna, J. – Ariyaeeinia, A. 2001. On the use of the Bayesian information criterion in multiple speaker detection. In: *Proceedings of Eurospeech 2001*, Scandinavia.

Stokoe, E. 2006. On ethnomethodology, feminism, and the analysis of categorial reference to gender in talk-in-interaction. *Sociological Review* 54: 467–94.

Tritschler, A. – Gopinath, R. 1999. Improved speaker segmentation and segments clustering using the bayesian information criterion. In: Proceedings of Eurospeech 1999. 679–682.

Trueba-Hornero B. 2008 Handling overlapped speech in speaker diarization. Master's thesis, Universitat Politecnica de Catalunya, May 2008.

Vandecatseye, A. – Martens, J.-P. 2003. A fast, accurate and stream-based speaker segmentation and clustering algorithm. In: Proceedings of Eurospeech 2003, Geneva, Switzerland, 941–944.

Vescovi, M. – Cettolo, M. – Rizzi, R. 2003. A DP algoritm for speaker change detection. In: *Proceedings of Eurospeech 2003*.

Wooters C. – Huijberts M. 2007. The ICSI RT07s speaker diarization system. In Proceedings of of the Rich Transcription 2007 Meeting Recognition Evaluation Workshop, 2007. Baltimore, MD.

Ying, D. – Yan, Y. – Dang, J. – Soong, F. 2011. Voice Activity Detection Based On An Unsupervised Learning Framework. In: *IEEE Transactions on Audio, Speech and Language Processing* 19/8: 2624–2633.

7. AZ ÉRTEKEZÉS TÉMAKÖRÉBEN MEGJELENT PUBLIKÁCIÓK

Beke A, Gósy M 2014. Phonetic analysis and automatic prediction of vowel duration in Hungarian spontaneous speech. INTERNATIONAL JOURNAL OF INTELLIGENT DECISION TECHNOLOGIES 10: 57-66.

Váradi V, Beke A 2013. Az artikulációs tempó variabilitása a felolvasásban. BESZÉDKUTATÁS 21: 26–42.

Szaszák Gy, Beke A. 2013. Using phonological phrase segmentation to improve automatic keyword spotting for the highly agglutinating Hungarian language. In: 14th Annual Conference of the International Speech Communication Association. Lyon, Franciaország, 2013.08.25–2013.08.29.

Neuberger T, Beke A 2013. Automatic laughter detection in spontaneous speech using GMM-SVM method. In: Habernal I, Matousek V (szerk.) Text, Speech, and Dialogue: 16th International conference, TSD 2013, Pilsen, Czech Republic, September 1–5, 2013. Proceedings. Berlin; H. eidelberg: Springer Verlag, 2013. 113–120.

Gósy M, Bóna J, Beke A, Horváth V 2013. A kitöltött szünetek fonetikai sajátosságai az életkor függvényében. BESZÉDKUTATÁS 21: 121–143.

Beke A, Szaszák Gy, Váradi V 2013. Automatic phrase segmentation and clustering in spontaneous speech In: IEEE 4th International Conference on Cognitive Infocommunications, CogInfoCom 2013, December 2–5, 2013

Szaszák, György and Beke, András: Exploiting Prosody for Syntactic Analysis in Automatic Speech Understanding, Journal of Language Modelling, 0(1) 143–172. (2012)

Szaszák Gy, Beke A 2012. Statisztikai módszerek alkalmazása a szintaktikai szerkezet és a beszédjel prozódiai szerkezetének feltérképezéséhez olvasott és spontán beszédben In: Gósy M (szerk.) Beszéd, adatbázis, kutatások. Budapest: Akadémiai Kiadó, 2012. 236–250.

Szaszák Gy, Beke A 2012. Automatic prosodic and syntactic analysis from speech in Cognitive Infocommunication. In: IEEE (szerk.)3rd IEEE International Conference on Cognitive Infocommunications. CogInfoCom 2012. Proceedings. Kosice, Szlovákia, 2012.12.02–2012.12.05.

Gósy M, Gyarmathy D, Horváth V, Gráczi TE, Beke A, Neuberger T, Nikléczy P 2012. BEA: Beszélt nyelvi adatbázis. In: Gósy M (szerk.) Beszéd, adatbázis, kutatások. Budapest: Akadémiai Kiadó, 2012. 9–24.

Beke A, Szaszák Gy 2012. Unsupervised clustering of prosodic patterns in spontaneous speech In: Sojka P, Horák A, Kopeček I, Pala K (szerk.) Text, Speech and Dialogue: 15th International Conference, TSD 2012, Brno, Czech Republic, September 3–7, 2012. Proceedings. Berlin: Springer, 2012. 648–655.

Beke A, Gósy M, Horváth V 2012. Gyakorisági vizsgálatok spontán beszédben. BESZÉDKUTATÁS 20: 260–277.

Beke A, Gósy M 2012. Characteristic and spectral features used in automatic prediction of vowel duration in spontaneous speech. In: IEEE (szerk.) 3rd IEEE International Conference on Cognitive Infocommunications. CogInfoCom 2012. Proceedings. Kosice, Szlovákia, 2012.12.02–2012.12.05.

Beke A 2012. Beszélőfelismerés kevert Gauss–modellekkel. In: Markó Alexandra (szerk.) Beszédtudomány: Az anyanyelv–elsajátítástól a zöngekezdési időig. Budapest: ELTE és MTA Nyelvtudományi Intézete, 2012. 335–352.

Beke A 2012. Beszéddetektálás spontán beszédben a beszélőváltás–detektáláshoz. In: Váradi T (szerk.) VI. Alkalmazott Nyelvészeti Doktoranduszkonferencia: Budapest, 2012. 02. 03. Budapest: MTA Nyelvtudományi Intézet, 2012. 14–23.

Beke A 2012. Az egyszerre beszélések automatikus osztályozása spontán magyar társalgásokban. In: Bárdosi Vilmos (szerk.) Tanulmányok: Nyelvtudományi Doktori Iskola, Budapest: ELTE BTK, 2012. 23–39.

Szaszák Gy, Nagy K, Beke A 2011. Analysing the correspondence between automatic prosodic segmentation and syntactic structure. In: Piero Cosi, Renato De Mori, Giuseppe Di Fabbrizio, Roberto Pieraccini (szerk.) Interspeech 2011, 12th Annual Conference of the International Speech Communication Association. Firenze, Olaszország, 2011.08.27–2011.08.31.

Gósy M, Beke A, Horváth V 2011. Temporális variabilitás a spontán beszédben. BESZÉDKUTATÁS 19: 5–30.

Beke A 2011. Szókezdetek automatikus osztályozása spontán beszédben. MAGYAR NYELVŐR 135: 226–241.

Beke A, Szaszák Gy 2010. Szótagok automatikus osztályozása spontán beszédben spektrális és prozódiai jellemzők alapján. In: Tanács Attila, Vincze Veronika (szerk.) VII. Magyar Számítógépes Nyelvészeti Konferencia: MSZNY 2010, Szeged, Szegedi Tudományegyetem, 2010. 236–249.

Beke A, Szaszák Gy 2010. Kísérlet a szintaktikai szerkezet részleges automatikus feltérképezésére a prozódiai szerkezet alapján. In: Tanács Attila, Vincze Veronika (szerk.) VII. Magyar Számítógépes Nyelvészeti Konferencia: MSZNY 2010, Szeged, Szegedi Tudományegyetem, 2010. 178–190.

Beke A, Szaszák Gy 2010. Automatic recognition of schwa variants in spontanneous Hungarian speech. ACTA LINGUISTICA HUNGARICA 57:(2–3) 329–353.

Beke A 2009. A beszélő hangtartományának vizsgálata: Néhány statisztikai jellemző az alapfrekvencia–eloszlásról. In: Keszler Borbála, Tátrai Szilárd (szerk.) Diskurzus a grammatikában – grammatika a diskurzusban Budapest: Tinta Könyvkiadó, 2009. 83–91.

Beke A 2008. Az alapfrekvencia–eloszlás modellezése a beszélőfelismeréshez. ALKALMAZOTT NYELVTUDOMÁNY 8:(1–2) 121–133.

Beke A 2008. A felolvasás és a spontán beszéd alaphangszerkezeteinek vizsgálata. BESZÉDKUTATÁS 16: 93–107.

8. AZ ÉRTEKEZÉS TÉMAKÖRÉBEN TARTOTT ELŐADÁSOK

2007. április: A kérdés–válasz prozódiája számítógépes vizsgálattal. OTDK, Székesfehérvár.
2007. április: A kérdés–válasz dallamszerkezetének fonetikai vizsgálata magyar nyelvű társalgásokban. FÉLÚTON Konferencia, Budapest.
2008. április: A beszélő személy felismerése: az automatikus formánslekérdezés eredményei. FÉLÚTON Konferencia, Budapest.
2008. november: A beszélő hangtartományának vizsgálata (néhány statisztikai jellemző az alapfrekvencia–eloszlásról).: Diskurzus a grammatikában – grammatika a diskurzusban (Új nézőpontok a magyar nyelv leírásában 2.) Konferencia, Budapest, 2008. nov. 11–12.
2009. november: A svávariációk automatikus felismerése magyar nyelvű spontán beszédben. Beszédkutatás 2009. Budapest. 2009. (társszerzőségben:Szaszák György)
2010. november 23. Temporális variabilitás a spontán beszédben. Kultúra és nyelv, kulturális nyelvészet – Új nézőpontok a magyar nyelv leírásában 3. ELTE BTK: Budapest (társszerzőségben: Horváth Viktória és Gósy Mária).
2010. december 1. Szótagok automatikus osztályozása spontán beszédben spektrális és prozódiai jellemzők alapján. VII. Magyar Számítógépes Nyelvészeti Konferencia (társszerzőségben: Szaszák György)
2011. május 22.–26. Figyi, ki beszél most? A beszélők automatikus osztályozása a spontán társalgásokban. XIII. Balatonalmádi Pszicholingvisztikai Nyári Egyetem.
2011. május 22.–26. A hezitációs jelenségek gépi osztályozása a spontán beszédben. XIII. Balatonalmádi Pszicholingvisztikai Nyári Egyetem. (társszerzőségben: Horváth Viktóriával)
2011. agugusztus 28–31. Analysing the correspondence between automatic prosodic segmentation and syntactic structure. Interspeech 2011, Firenze, Olaszország (társszerzőségben: Szaszák Györggyel)
2011. október 27–28. Gyakorisági mutatók a spontán beszédben. Beszédkutatás konferencia. (társszerzőségben: Gósy Máriával és Horváth Viktóriával)
2011. október 27–28. Kísérlet a szintaktikai szerkezet részleges automatikus feltérképezésére a prozódiai szerkezet alapján. Beszédkutatás konferencia. (társszerzőségben: Szaszák Györggyel)
2011. október 27–28. Az ismétlések automatikus osztályozása a spontán beszédben. Beszédkutatás konferencia. (társszerzőségben: Gyarmathy Dorottyával)
2011. november 17. Beszélők szegmentálása és osztályozása társalgásban. A Magyar Tudomány Ünnepe 2011 Beszédadatbázisok a kutatásban és az alkalmazásban. MTA NYTUD. Budapest.
2011. december 1–2. A szintakai szerkezet automatikus feltérképezése a beszédjel prozódiai elemzése alapján. VIII. Magyar Számítógépes Nyelvészeti Konferencia (társszerzőségben: Szaszák György)

2012. március: Toward Exploring the Prosodic Structure of Spontaneous Speech by Focusing on Automatic Modelling, IAST Workshop, Dublin (társszerzőségben: Szaszák György)
2012. április: A beszélő személy gépi felismerése. Fonetikanap, ELTE BTK, Budapest.
2012. szeptember: Unsupervised Clustering of Prosodic Patterns in Spontaneous Speech. TSD, Brno, (társszerzőségben: Szaszák György)
2012. szeptember: A szintaxis és a prozódia kapcsolata, BEA Workshop, MTA Nyelvtudományi Intézet, Budapest (társszerzőségben: Szaszák György)
2012: december: Automatic prosodic and syntactic analysis from speech in Cognitive Infocommunication, CogInfoCom konferencia, Kassa (társszerzőségben: Szaszák György)
2012: december: Characteristic and spectral features used in automatic prediction of vowel duration in spontaneous speech, CogInfoCom konferencia, Kassa (társszerzőségben: Gósy Mária)
2013. március: Automatic identification of discourse markers in spontaneous speech for speaker diarization. SJUSK 2012. Copenhagen
2013. március: Automatic laughter detection in Hungarian spontaneous speech using GMM/ANN hybrid method. SJUSK 2012. Copenhagen (társszerzőségben: Neuberger Tilda)
2013. március: Automatic classification of repeated words in Hungarian spontaneous speech. ExAPP 2013. Copenhagen (társszerzőségben: Gyarmaty Dorottya)
2013. augusztus 25–29. Using Phonological Phrase Segmentation to Improve Automatic Keyword Spotting for the Highly Agglutinating Hungarian Language. INTERSPEECH 2013, Lyon, Farnciaország (társszerzőségben: Szaszák Györggyel)
2013. december 7–9. Temporal variability in spontaneous Hungarian speech. 6th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (társzerzőségben: Gósy Mária, Horáth Viktória)
2013. szeptember 1–5. Automatic Laughter Detection in Spontaneous Speech Using GMM–SVM Method.Text, Speech, and Dialogue – 16th International Conference, TSD 2013, Pilsen, Czech Republic (társzerzőségben: Neuberger Tilda)
2013. szeptember 1–5. A Logistic Regression Approach for the Improvement of Keyword Spotting based on Phonological Phrasing. Text, Speech, and Dialogue – 16th International Conference, TSD 2013, Pilsen, Czech Republic (társzerzőségben: Szaszák György)